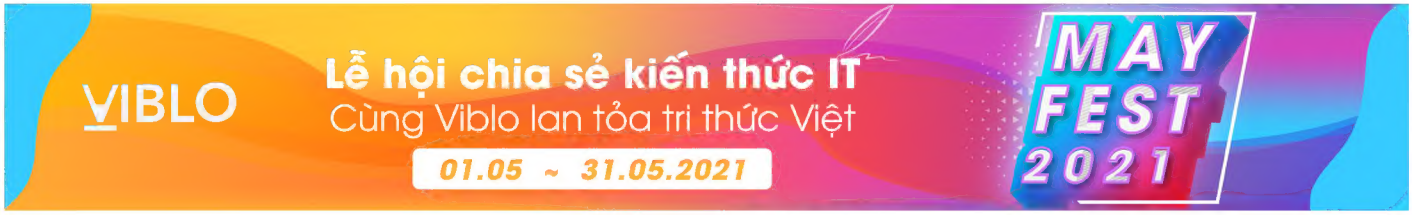


Lê Ngọc Sơn @uzumaki2205 Follow
215 7 10

Published Tuesday, 1:00 PM - 7 min read

141 0 0

Telegram Bot - Cào Dữ Liệu Từ VnExpress Bằng Python

Chào mọi người, sau bao ngày với các bài viết về lỗi bảo mật thì hôm nay mình sẽ đổi gió tí nhỉ :v

Vì thế nên hôm nay mình sẽ hướng dẫn mọi người làm 1 con bot Telegram bằng Python nhé.

Nếu mà nhắc tới Telegram thì có lẽ đã không còn xa lạ gì đối với mọi người bởi tính tiện dụng trong công việc của nó, nhưng có bao giờ bạn nghĩ rằng sẽ biến tele của bạn thành 1 con bot dùng để lấy dữ liệu nhanh chóng hay chưa ??

Một phần bởi vì Telegram đã public API cho mọi người sử dụng, nhưng nếu mà free thì tội gì mình không thử 😆

Vậy tại sao lại là Telegram

- Thứ nhất: Tất nhiên là vì tính riêng tư của tele được người dùng ưa chuộng hơn so với Facebook hoặc các mạng xã hội khác.
- Thứ hai: Là vì Telegram rất nhẹ, gửi tin nhắn nhanh, icon, gif đẹp mắt, blabla, ... (chứ ai đời lại như ứng dụng nào đó được đánh giá 3.5 sao trên Store :v)

OK kể chuyện bấy nhiu đủ rồi, chúng ta cùng bắt đầu thôi!!

Giới Thiệu

Nếu bạn đi làm trong 1 công ty hay doanh nghiệp, đa phần người ta sẽ yêu cầu bạn dùng những ứng dụng yêu cầu mang tính riêng tư 1 tí (tất nhiên không phải facebook rồi) để hạn chế khả năng bị rò rĩ dữ liệu công ty ra bên ngoài như Zalo hoặc Telegram, etc....

Tất nhiên 1 phần là tại vì lượng người truy cập mạng xã hội quá đông sẽ đến đến tình trạng không an toàn về mặt dữ liệu (Rõ ràng mọi người đã nghe đến hack facebook, hack blabla, không phải là không có nhưng rất ít khi bao giờ nghe đến hack telegram hay hack zalo nhiều đâu đúng không 😃)

Bởi vì món mồi béo bở của hacker là nơi có nhiều người truy cập và nơi có lượng lớn người dùng giao dịch để dễ dàng thực hiện hành vi của mình.

Nhưng mình ở đây không phải pr cho Telegram mà là hướng dẫn mọi người tạo chat bot ạ :v Mọi người có thể tự dùng và cảm nhận nhé.

1. Quy Trình Tạo ChatBot Trên Telegram

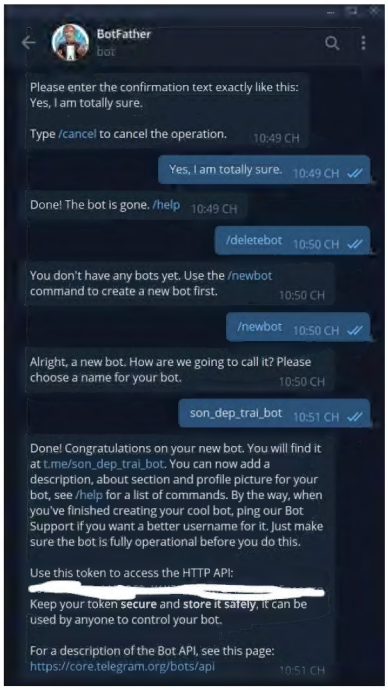
1.1 Bước 1: Nhờ BotFather

(Lệnh /delete các bạn đừng để ý nhé)

Chúng ta có thể tạo bot bằng cách tìm **BotFather** trên thanh search rồi chat với **BotFather** là *"con muốn tạo bot ạ :v"* bằng lệnh `/newbot`

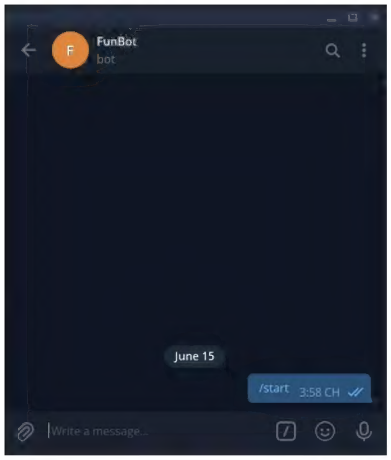
Sau đó BotFather sẽ hỏi rằng ta muốn đặt tên bot là gì thì chúng ta cứ bấm vào thôi.

Sau khi tạo xong Bot thì BotFather sẽ gửi lại 1 **API KEY** mà mình đã xóa trong hình đi, chúng ta sẽ dùng **Key** này để tạo kết nối với Bot



1.2 Bước 2: Chat Thử Với Bot

Chúng ta sẽ chat thử với Bot xem đã ok chưa, tất nhiên là Bot chưa thể tự trả lời rồi :v



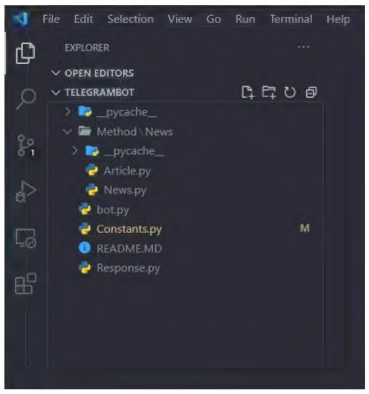
2. Bắt Đầu Code

2.1 Chuẩn Bị

- Python (tốt nhất là version 2.7 trở lên, mình thì đang dùng 3.8)

Sau khi cài đặt Python chúng ta sẽ tạo cấu trúc của Project

- Folder Method dùng để chứa các Function xử lý rồi trả về kết quả cho Bot
    - Ở đây mình đã tạo sẵn 1 Folder **News** để chứa 2 File là **Article.py** và **News.py** dùng để xử lý cào dữ liệu từ VnExpress về.
- File **Constants.py** dùng để chứa các giá trị không đổi như **API KEY**
- File **Response.py** dùng để chứa các câu trả lời khi gặp các câu hỏi đơn giản ví dụ như hi, helo, chào, ...
- File **bot.py** sẽ là File xử lý chính trong chương trình.

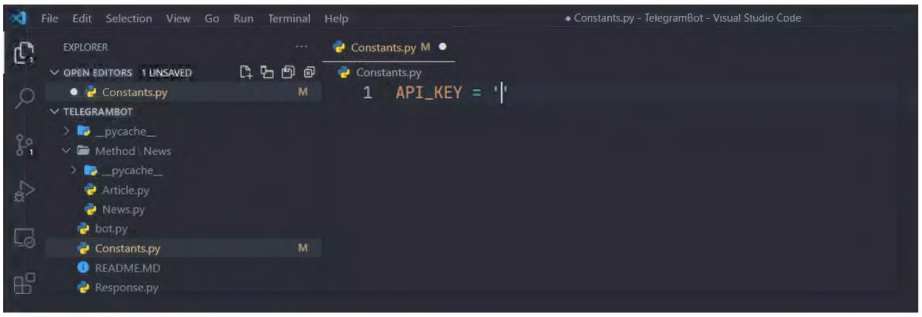


Ngoài cách mà mình sắp hướng dẫn dưới dưới đây, còn 1 cách đơn giản hơn đó là dùng request để gửi tin nhắn thông qua url thôi, các bạn có thể tham khảo tại đây nhé.

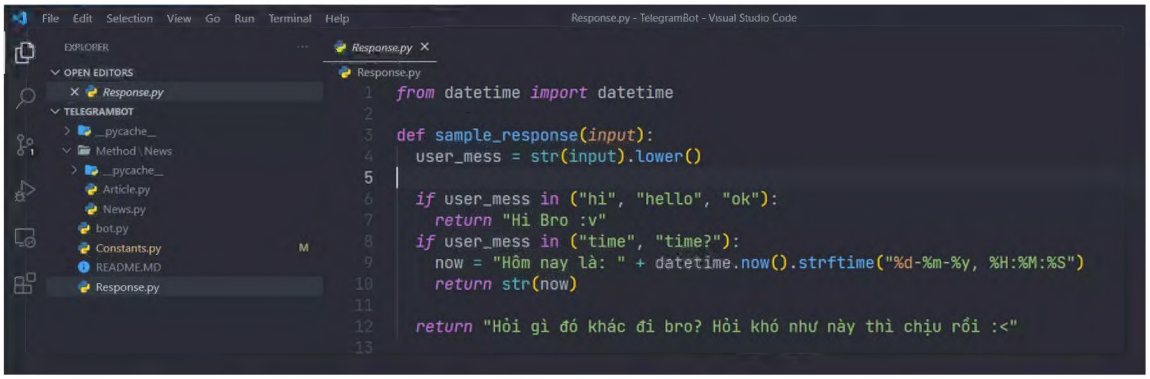
2.2 Bắt Đầu Code

Bước 1: Code Cấu Trúc

Vào file **Constants.py** vào thêm biến **API_KEY** và đặt key lúc nãy nhận được từ BotFather vào.



Sang file **Response.py** và thêm câu hỏi vào




```
from datetime import datetime

def sample_response(input):
  user_mess = str(input).lower()

  if user_mess in ("hi", "hello", "ok"):
    return "Hi Bro :v"
  if user_mess in ("time", "time?"):
    now = "Hôm nay là: " + datetime.now().strftime("%d-%m-%y, %H:%M:%S")
    return str(now)
  // Có thể thêm các câu hỏi mà người dùng nhập vào ở đây

  return "Hỏi gì đó khác đi bro? Hỏi khó như này thì chịu rồi :<"
```

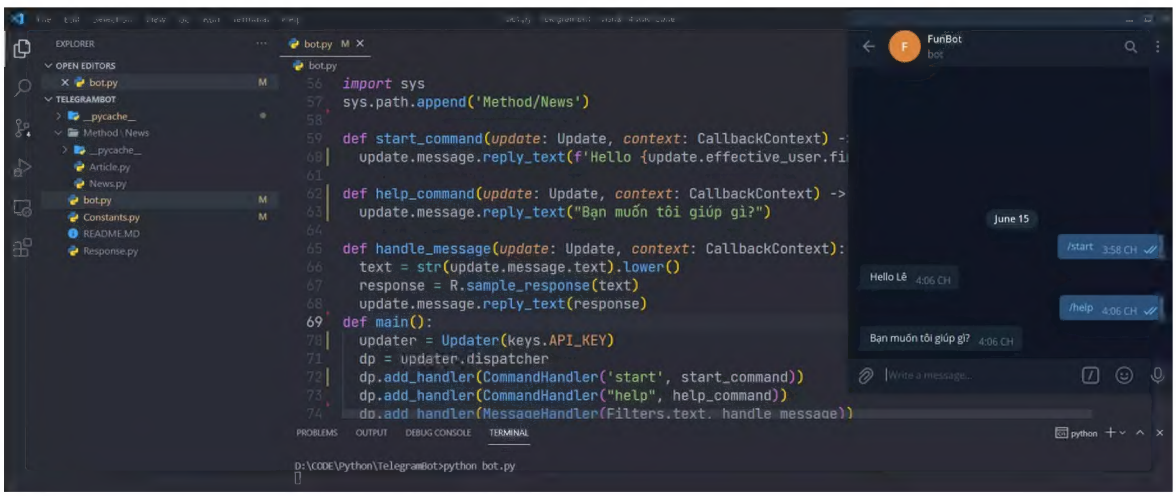
Sang file bot.py chúng ta sẽ cài các package sau vào để dễ dàng cho việc code

```
pip install requests # Thư viện dùng để tạo các request
pip install beautifulsoup4 # Thư viện để bóc tách html
pip install python-telegram-bot
```

```
import Constants as keys
import Response as R
from telegram import Update
from telegram.ext import Updater, CommandHandler, MessageHandler, CallbackContext, Filters
import json
import sys
sys.path.append('Method/News') # Import File News.py trong đường dẫn Method/News vào

# Hàm khởi động với lệnh /start trong telegram
def start_command(update: Update, context: CallbackContext) -> None:
  update.message.reply_text(f'Hello {update.effective_user.first_name}')

# Hàm khởi động với lệnh /help trong telegram
def help_command(update: Update, context: CallbackContext) -> None:
  update.message.reply_text("Bạn muốn tôi giúp gì?")

# Hàm khởi động với các câu hỏi trong file Response.py
def handle_message(update: Update, context: CallbackContext):
  text = str(update.message.text).lower()
  response = R.sample_response(text)
  update.message.reply_text(response)

def main():
  updater = Updater(keys.API_KEY)
  dp = updater.dispatcher

  # Define các lệnh mà mọi người cần
  # Param thứ nhất là nội dung câu lệnh, param thứ 2 là hàm để chạy lệnh đó
  dp.add_handler(CommandHandler('start', start_command))
  dp.add_handler(CommandHandler("help", help_command))
  dp.add_handler(MessageHandler(Filters.text, handle_message))
  updater.start_polling()
  updater.idle()

main()
```

CommandHandler là cơ chế đưa vào câu lệnh trong Telegram, còn **MessageHandler** sẽ xử lý các input thông thường.

Thử chạy chương trình và chat với Bot



Bước 2: Code cào dữ liệu từ VnExpress

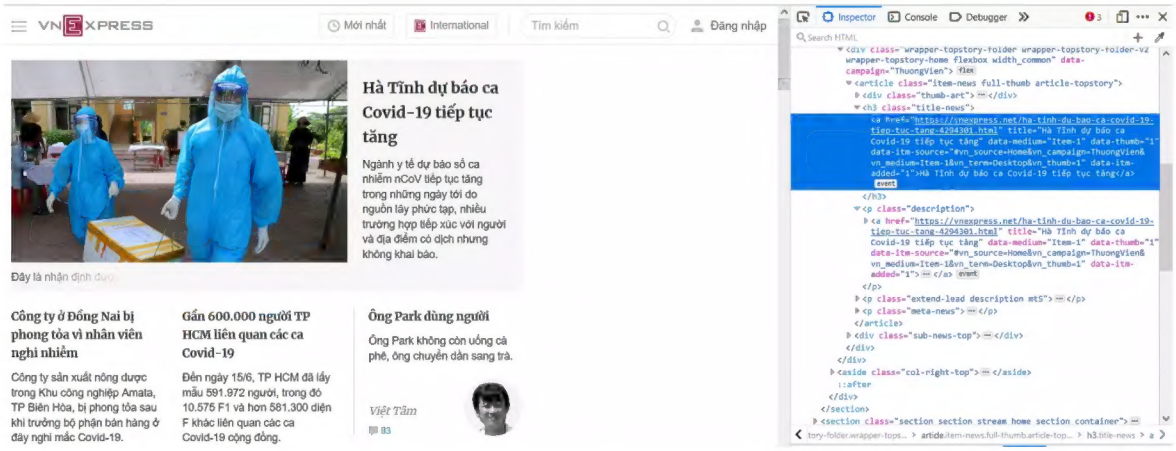
Mở file **Article.py** và định nghĩa **class Article** (mình sẽ sử dụng phương pháp lập trình hướng đối tượng ứng dụng vào)

Chúng ta sẽ lấy tiêu đề, link và description của bài viết (không cần lấy ảnh, vì khi gửi link ảnh sẽ tự hiện lên thôi)

```
class Article:
    def __init__(self, title, link, description):
        self.title = title
        self.link = link
        self.description = description
```

Chúng ta sẽ phân tích VnExpress 1 chút

Ta sẽ lấy đoạn từ trong thẻ **<article>**, mình để ý ở phần thẻ **p class=description** cũng có gần hầu hết tiêu đề, link và description mình cần NHƯNG mình vẫn sẽ lấy từ phần article, nhìn cho nó chuyên nghiệp ✌️



- File **News.py**

```
import requests
from bs4 import BeautifulSoup
import json
from Article import Article

baseUrl = "https://vnexpress.net/"

def GetNews(limit_news = 20): # Tạo biến limit_news để lấy số lượng tin mà mình cần thôi
    s = requests.Session() # Store sesstion lại
    response = s.get(baseUrl) # Thực hiện Get request
    soup = BeautifulSoup(response.content, 'html.parser') # Đưa vào biến soup chuẩn bị bóc tách dữ liệu
    article = soup.select("article.item-news", limit=limit_news) # Tách dữ liệu phần thẻ article ra

    listArticle = []
    for element in article:
        title = element.select("h3.title-news > a") # Lấy phần thẻ chứa title
        description = element.select("p.description > a") # Lấy phần thẻ chứa description
        for x in range(len(title)): # serialize object này lại thành json để lấy dữ liệu dễ dàng hơn
            listArticle.append(json.dumps(Article(title[x]['title'], title[x]['href'], description[x].te
    return listArticle
```

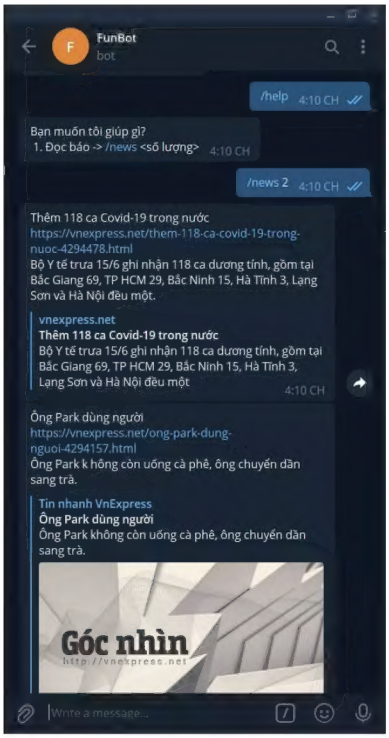
Mình đã test cào xuống dữ liệu ổn rồi nhé

Tiếp tục trở lại file **bot.py** chỉnh sửa thêm các lệnh mới

Full Code:

```
import Constants as keys
import Response as R
from telegram import Update
from telegram.ext import Updater, CommandHandler, MessageHandler, CallbackContext, Filters
import json
import sys
sys.path.append('Method/News')
import News

print("Bot Starting....")

def start_command(update: Update, context: CallbackContext) -> None:
  update.message.reply_text(f'Chào {update.effective_user.first_name}')

def help_command(update: Update, context: CallbackContext):
  update.message.reply_text("Bạn muốn tôi giúp gì? \n 1. Đọc báo -> /news <số lượng>")

def news_command(update: Update, context: CallbackContext):
  try:
    limit_news = int(context.args[0]) # Lấy tham số từ input truyền vào -> cào về bao nhiêu tin
    news = News.GetNews(limit_news)
    for x in range(0, len(news)): # Deserialize dữ liệu json trả về từ file News.py lúc nãy
      message = json.loads(news[x])
      update.message.reply_text(message['title'] + "\n"
        + message['link'] + "\n" + message['description'])
  except (IndexError, ValueError):
    update.message.reply_text('Vui lòng chọn số lượng tin hiển thị!!')

def handle_message(update: Update, context: CallbackContext):
  text = str(update.message.text).lower()
  response = R.sample_response(text)
  update.message.reply_text(response)

 # Function dùng để xác định lỗi gì khi có thông báo lỗi
def error(update: Update, context: CallbackContext):
  print(f"Update {update} cause error {context.error}")

def main():
  updater = Updater(keys.API_KEY, use_context=True)
  dp = updater.dispatcher
  dp.add_handler(CommandHandler("start", start_command))
  dp.add_handler(CommandHandler("help", help_command))
  dp.add_handler(CommandHandler("news", news_command))
  dp.add_handler(MessageHandler(Filters.text, handle_message))
  dp.add_error_handler(error)

  updater.start_polling()
  updater.idle()

main()
```

Ok test thử nào 😃



Chỉ cần nhập số lượng tin là Bot sẽ hiểu là chỉ được phép lấy bấy nhiu tin trả về

Ngoài ra, mọi người có thể thêm một số hàm rồi gọi sang file bot.py rất linh hoạt, không cần phải chỉnh sửa nhiều.

Tổng Kết

Nói chung thì code cũng đơn giản đúng không mọi người, chỉ 1 vài dòng code là chúng ta đã có 1 con Bot xịn xò rồi.

Telegram API còn cung cấp cho chúng ta không chỉ gửi text thôi mà còn là hình ảnh, markdown, game nữa, các bạn có thể tìm hiểu thêm nhé.

Không chỉ dừng với việc cào dữ liệu, chúng ta còn thể tích hợp thêm AI vào bằng cách cài 1 số thư viện như tensorflow phân tích ảnh các kiểu :v

Vậy là phần hướng dẫn của mình đến đây là kết thúc, cảm ơn mọi người đã đọc hết bài viết. Nếu bài viết có nhiều chỗ chưa hiểu hay sơ sót, mọi người có thể comment lại để mình có thể trả lời và giải đáp thắc mắc nhé.

Chúc mọi người thành công!!

Tham Khảo

- BeautifulSoup4 Document: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- Requests Document: https://docs.python-requests.org/en/master/
- Python Telegram Bot Document: https://python-telegram-bot.readthedocs.io/en/stable/

Advance Python | bot | Bot Chat | telegram | Telegram Messenger

All Rights Reserved

RESOURCES

Posts
Questions
Videos
Discussions
Tools
System Status
Organizations
Tags
Authors
Recommend System
Machine Learning

SERVICES

Viblo Code
Viblo CV
Viblo CTF
Viblo Learning

MOBILE APP

GET IT ON Google Play
Download on the App Store

LINKS

© 2021 **Viblo**. All rights reserved.

About Us Feedback Help FAQs RSS Terms DMCA PROTECTED